Statistical Analysis of proteomic data

Grier P Page Ph.D.
Assistant Professor
Section on Statistical Genetics
Department of Biostatistics
School of Public Health

Gpage@uab.edu
4-4930
Ryals 317D
Keeping Up with the Microarray Literature: How Many Can You Read Per Day?

Microarray Articles in PubMed

- All Microarray papers
- Statistical Microarray Papers

From Mehta, Tanik, & Allison.
A Perspective on Statistics

• We study:

• We wish to obtain knowledge about:

Samples

Populations

Data

Nature
Things Statisticians Do:

Develop Design & Analysis Procedures to Facilitate:

• **Measurement** – (e.g., produce a variable $Y'$ that represents $Y$).

• **Prediction** – (e.g., ‘impute’ unobserved values of $X$ using observed $Y$).

• **Estimation** – (e.g., estimate $\Delta = \mu_1 - \mu_2$).

• **Inference** – (e.g., conclude whether $\delta = 0$).

• **Classification** – (e.g., for $j = 1$ to $k$, sort the $Y_j$ into $m < k$ groups).
Epistemological Foundations

• Epistemology is the study of how we come to have and what constitutes knowledge.

• Given a set of statistical procedures judged to be valid, a sound epistemological foundation for biological science comes, in part, from the application of those procedures.

• But how do we derive knowledge about the validity of our statistical methods such that they are also enjoy a solid epistemological foundation?
Method Validation

Epistemologically Valid Frameworks: Induction & Deduction

- Deduction: i.e., mathematical proof.
- Induction:
  - Simulations
  - Plasmodes
- Composite Approaches: Application to multiple real data sets of unknown nature with methods of partially known properties.

A Circular & Epistemologically Invalid Framework

- Application to single real data sets of unknown nature.
What is High Dimensional Biology?

• High Dimensional Biology – is a broad topic covering biological systems where the number of variables is very large.

• Topics that often fall in HDB are microarray, proteomics, linkage, and genomics.

• HDB is also highly collaborative both ‘wet’ and ‘dry’ lab people.
Affymetrix type array
A. Nup42p affinity resin

Nup42p affinity resin

B. Nup49p affinity resin

Barnes class 02-04-03
What Do All These Topics Have in Common?

Lots and Lots and Lots of Numbers !!!
If you have numbers what do you do?

• **Statistics (and Design) !**

• Or as most of you think Statistics Ugh!

• Most of the statistics used in HDB are identical to statistical methods that have been used for years.

• The thought process that goes into design is also similar to those that have been used for years.
Design

• Design is the art of designing an experiment in such a way that the question that is being asked can be easily and unambiguously answered.

• The experimental hypothesis drives the design.
Statistics

• Methods for make inferences about a population as a whole by taking a sample.

• Statistics and design work in harmony with the biology, while design and statistical may be the cause of alterations in experiments, the biology is the *sine qua non*.
What are Statistics and Design?

- The goal of experimental design and statistical analysis is to allow an investigator to answer the question that they would like to ask correctly and efficiently.

- Often statisticians are a reality check. If you can’t explain your experiment to a statistician will it make sense in a publication?
Biological question
Differentially expressed genes
Sample class prediction etc.

Experimental design

Microarray experiment

Image analysis

Normalization

16-bit TIFF files

(Rfg, Rbg), (Gfg, Gbg), Signal

R, G, S

Estimation
Testing

Clustering

Discrimination

Biological verification and interpretation

From T. Speed
Quality Issues - I

• Known sources of non-biological error (not exhaustive) that must be addressed
  – Technician
  – Chip lot
  – Reagent/gel lot
  – Printer tip
  – Time of printing
  – Date
  – Fluidics well/ Scanner/ position on scanner
  – Order of scanning
  – Location
  – Cage/ Field position
  – Far and away the largest issue is labeling
Cluster Analysis of GG/BG Study

From Susan Hilsenbeck with permission
Quality Issues – II

• How to address these issues
  – Make the experiment as uniform as possible
    • Agree on exactly what defines the tissue to be used, use same technician, same chip lot, same reagents (always buy a little too much), same scanner, do sample extraction, labeling and hybridization on one day if possible, establish quality control
  – Randomize when uniformity is not possible
    • Don’t do all of condition 1 on day 1 and condition 2 on day 2
    • Randomize the time a chips sits waiting to be scanned
    • Randomize animal cage/plant field position
• Microarrays generate such a huge volume of data that is is possible to detect these issues, I suspect that northerns, Southern, RT-PCR, westerns, and more have similar problems.
Elements of Statistics

- Power – the probability of detecting something if it is there. Usually a function of sample size and size of difference to be detected
- Image Analysis
- Quality Control- normalization/transformation
- Normalization
- Statistical Analysis
  - Class discrimination
  - Class prediction
  - Class differentiation
- Annotation
- Bioinformatics issue
Image Analysis

• How do you go from an image to a number?
From Helen Kim
Which Size Circle?

Parts of other Proteins

Which Size Circle?

From Helen Kim
Image Analysis

Inside the boundary is spot (foreground), outside is not.

From T. Speed
Quality Control/Normalization

• Not all gels, chips, sequencing runs, etc are perfect
• Some are so bad they should be dropped
• Other can be fixed
  – Identify problem values/areas
  – Fix them – adjustments and normalization
Spatial plots: background from the two slides

From T. Speed
Liver

85IWHHTFYNELR^95

Kidney

197GYSFTTTAER^206

239SYELPDGQVITIGNER^254

Steve Barnes 2-11-03
Aslan et al., JBC 278:4194
Time of printing effects

Green channel intensities ($\log_2 G$). Printing over 4.5 days. The previous slide depicts a slide from this print run. From T. Speed/H Yang
Mean Normalization
Intensity correction only

Before

After
Composite normalization

Before and after composite normalization

- MSP lowess curve
- Global lowess curve
- Composite lowess curve

(Other colours control spots)

From T. Speed
Statistical Analysis

- Statistical Analysis
  - Class discrimination
  - Class prediction
  - Class differentiation
Suppose we conduct a t-test of the difference between two means and obtain a p-value < .05. Does this mean:

a) There is less than a 5% chance that the results are due to chance.

b) If there really is no difference between the population means, there is less than a 5% chance of obtaining a difference this large or larger.

c) There is a 95% chance that if the study is repeated, the result will be replicated.

d) There is a 95% chance that there is a real difference between the two population means.

From H Kim
Class Discovery

• Data visualization
• Cluster analysis
  – Clustering
  – Self organizing maps
• Multidimensional scaling
• Similarity searching
Clustering

• There are a large number of clustering algorithms.
  – Hierarchical
  – Non-hierarchical
  – Different weights
  – All will give different answers.
  – None are statistical tests

From *Nature*
From Nature
K-Means Clustering

Source Unknown
Class Prediction

• Discriminate Analysis
  – Build a predictive model for future data based upon previous data.
  – Each new sample is assigned the probability that it will fall into one of the classes.

• Assign new samples to one of several groups
  – e.g. is a new tumor adenoma or squamous cell carcinoma
<table>
<thead>
<tr>
<th>GC B-Like</th>
<th>Activated B-Like</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Nature</td>
<td></td>
</tr>
</tbody>
</table>
From *Nature*
Class Differentiation

- Supervised Analysis
- What genes are most different between two or more groups
“There are other experiments, however, which cannot easily be repeated very often; in such cases it is sometimes necessary to judge the certainty of the results from a very small sample, which itself affords the only indication of the variability.”

-- Student (1908)
### Types of Statistical Tests and Approaches

<table>
<thead>
<tr>
<th>Type of Dependent Data</th>
<th>Type of Independent Data</th>
<th>One Sample (focus usually on estimation)</th>
<th>Two Samples</th>
<th>Multiple Samples</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical (dichotomous)</td>
<td>Categorical</td>
<td>1 Estimate proportion (and confidence limits)</td>
<td>2 Chi-Square Test</td>
<td>3 McNemar Test</td>
<td>4 Chi Square Test</td>
</tr>
<tr>
<td>Continuous</td>
<td>Categorical</td>
<td>8 Estimate mean (and confidence limit)</td>
<td>9 Independent t-test</td>
<td>10 Paired t-test</td>
<td>11 Analysis of Variance</td>
</tr>
<tr>
<td>Right Censored (survival)</td>
<td>Continuous</td>
<td>15 Kaplan Meier Survival</td>
<td>16 Kaplan Meier Survival for both curves, with tests of difference by Wilcoxon or log-rank test</td>
<td>17 Very unusual</td>
<td>18 Kaplan-Meier Survival for each group, with tests by generalized Wilcoxon or Generalized Log Rank</td>
</tr>
<tr>
<td></td>
<td>Continuous</td>
<td>20 Proportional Hazards analysis</td>
<td>21 Proportional Hazards analysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

After G. Howard
What should I use for 2-group testing?

- Equal sample sizes?
  - Yes: Equal variances?
    - Yes: TTEV
    - No: TTUV
  - No: Equal variances?
    - Yes: TTEV
    - No: CC2

- Normality?
  - Yes: Equal variances?
    - Yes: TTEV
    - No: CC1
  - No: Equal variances?
    - Yes: CC7
    - No: CC1
Under the null hypothesis, the distribution of p-values is uniform on the interval $[0,1]$ regardless of the sample size and statistical test used (as long as that test is valid).

Under the alternative hypothesis, the distribution of p-values will tend to cluster closer to zero than to one.
Fitted mixture model to 12,625 P-values
### Testing Defined

<table>
<thead>
<tr>
<th>Conclusion</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>Null</td>
</tr>
<tr>
<td>Null</td>
<td>a</td>
</tr>
<tr>
<td>Alt</td>
<td>c</td>
</tr>
<tr>
<td>K-M</td>
<td>M</td>
</tr>
</tbody>
</table>

- $c =$ type 1 error (alpha) – false positive
- $b =$ type 2 error (beta) – false negative

\[
\text{FDR} = E \left( \frac{c}{c + d} \right)
\]
FDR - False Discovery Rate

• When many hypotheses are tested the sample size required for a Bonferroni corrected $p < 0.05$ were prohibitive in most contexts.

• Some attempts were made for intermediate adjustments
  – Lander and Botstein (1989) for linkage data

• Benjamini and Hochberg 1995 pulled together several streams of research on adjusting for multiple testing.
  – Developed method for setting an adjusted $p$-value that controlled for type I error
  – Like many statistical methods it has been ‘extended’ and abuse to a FDR estimating procedure

• Methods were developed for epidemiology and genetic studies, but were adapted for HDB studies
Under the null hypothesis, the distribution of p-values is uniform on the interval \([0,1]\) regardless of the sample size and statistical test used (as long as that test is valid).

Under the alternative hypothesis, the distribution of p-values will tend to cluster closer to zero than to one.
Family Wise Error Rate vs. False Discovery Rate

• Traditional FWER
  – Bonferroni $\alpha^* = \frac{\alpha}{n}$
  – Sidak $(1-(1-a)^n)$
    • Very conservative
    • Minimize False discovery rates
    • Assume independence

• False Discovery Rate
  – Designed to estimate the rate of error
Power and Sample Size

- This is where microarray experiments get the most criticism.
- Experiments performed without replication
- Impression that arrays much more expensive than they are now
- Belief that microarrays are not liable to the same experimental error that experiments are
- There also has not been a good way to calculate sample size
Power

• All power and sample size calculations require and estimate of population variability
• For microarrays we use a pilot project
• Based upon the posterior probability that a gene is differentially expressed it test statistic may be increased as a function of proposed increase in sample size
Power For Powerful Effect

Sample Size vs. Probability

- TP 0.1
- TP 0.05
- TP 0.01
- TP 0.001
- TP 0.0001
- Zeta 0.1
- Zeta 0.05
- Zeta 0.01
- Zeta 0.001
- Zeta 0.0001
- Zeta 0.00001
Data Interpretation

- The most time consuming portion of a HDB experiment is the interpretation.
- Many databases and resources exist.
  - Dr. Loraine talked about these in great detail.
**a posteriori vs. a Priori data interpretation**

- Many people get the data and then stare at it and tell a story based on their subjective observations about the data.
- *A posteriori* observations are highly biased.
- *A priori* observations require knowledge of pathway, gene family, etc. There can be a large number of classes.
Global/Meta Analytical Tests of Pathways

**Premise:** We can learn something additional and/or test with more power if we consider the fact that genes may exist within ‘families.’ Several Tests –

- Fisher’s meta analytical tests – combine the individual p-values from n genes $\sim \chi^2_{(2n-2)}$
- Vote Counting methods
  - Onto-express
  - GSEA
- Normalize all the data to Z scores and compare the expression levels
- Issues even under $H_0$ if genes in a pathway are correlated there will be an increase in type 1 error
- Address FEWR vs FDR per group
Gene Family-Based Hypothesis Testing: 

What people say they are testing vs what they are testing.

Which Null?
1. None of the genes in family c are differentially expressed.
2. The proportion of genes in family c that are differentially expressed is equal to the proportion of genes in the remainder of the genome that are differentially expressed.
3. The correlation matrix among the expression levels of the genes in family c is an identity matrix.
4. The correlation matrix among the expression levels of the genes in family c is the same across experimental conditions.
5. The intersection of #1 and #3.

Mootha et al (2003). “We introduce an analytical strategy, Gene Set Enrichment Analysis, designed to detect modest but coordinate changes in the expression of groups of functionally related genes.”

This implies that the null of interest is #1, but the test appears to be the intersection of #2 and #3.
Global/Meta Analysis

<table>
<thead>
<tr>
<th>Function Name</th>
<th>Total</th>
<th>P-Value</th>
<th>FDR</th>
<th>Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>inflammatory response</td>
<td>71</td>
<td>1.11E-16</td>
<td>4.72E-14</td>
<td>4.72E-14</td>
</tr>
<tr>
<td>immune response</td>
<td>95</td>
<td>8.44E-15</td>
<td>1.79E-12</td>
<td>3.59E-12</td>
</tr>
<tr>
<td>epidermal differentiation</td>
<td>38</td>
<td>1.65E-11</td>
<td>2.34E-09</td>
<td>7.02E-09</td>
</tr>
<tr>
<td>cell-cell signaling</td>
<td>100</td>
<td>3.14E-10</td>
<td>3.34E-08</td>
<td>1.34E-07</td>
</tr>
<tr>
<td>cell adhesion</td>
<td>77</td>
<td>5.72E-09</td>
<td>4.86E-07</td>
<td>2.43E-06</td>
</tr>
<tr>
<td>chemotaxis</td>
<td>43</td>
<td>8.73E-09</td>
<td>6.18E-07</td>
<td>3.71E-06</td>
</tr>
<tr>
<td>cellular defense response</td>
<td>40</td>
<td>1.74E-08</td>
<td>1.06E-06</td>
<td>7.39E-06</td>
</tr>
<tr>
<td>development</td>
<td>80</td>
<td>3.44E-08</td>
<td>1.83E-06</td>
<td>1.46E-05</td>
</tr>
<tr>
<td>antimicrobial humoral response</td>
<td>45</td>
<td>9.90E-08</td>
<td>4.68E-06</td>
<td>4.21E-05</td>
</tr>
<tr>
<td>response to viruses</td>
<td>18</td>
<td>7.16E-07</td>
<td>3.04E-05</td>
<td>3.04E-04</td>
</tr>
<tr>
<td>cell surface receptor linked signal transduction</td>
<td>54</td>
<td>3.29E-06</td>
<td>1.27E-04</td>
<td>1.40E-03</td>
</tr>
<tr>
<td>cell motility</td>
<td>47</td>
<td>3.55E-06</td>
<td>1.26E-04</td>
<td>1.51E-03</td>
</tr>
<tr>
<td>cell proliferation</td>
<td>79</td>
<td>1.81E-05</td>
<td>5.90E-04</td>
<td>7.67E-03</td>
</tr>
<tr>
<td>protein biosynthesis</td>
<td>6</td>
<td>1.81E-05</td>
<td>5.49E-04</td>
<td>7.69E-03</td>
</tr>
<tr>
<td>skeletal development</td>
<td>36</td>
<td>2.59E-05</td>
<td>7.34E-04</td>
<td>1.10E-02</td>
</tr>
</tbody>
</table>
Qualities of the Tazmanian Devil, Wile E. Coyote, and Elmer Fudd

**Tazmanian Devil**
- Australian (Tasmania)
- had his own cartoon show, Tasmanian
- first appeared in early '30s
- spits and slathers uncontrollably
- often distracted by the devils
- not very intelligent
- speech impediment
- preferred game is rabbit
- usually hunts in forests
- straightforward approach

**Wile E. Coyote**
- brown fur
- fast
- personified (human-like)
- wearing overalls
- preferred game is road runner
- usually hunts in deserts
- plans elaborately
- buys only ACME brand
- genius
- uses weapons to hunt
- co-created by Chuck Jones
- Who Framed Roger Rabbit cameo

**Elmer Fudd**
- has no hair
- uses a false for Bugs Bunny
- human
- won an Academy Award in 1940

“Transcription alterations in WS were strikingly similar to those in normal aging: 91% of annotated genes displayed similar expression changes in WS and in normal aging, 3% were unique to WS, and 6% were unique to normal aging. “

Yet, by chance alone, (A-B) will generally be correlated with (A-C). Simulating their data as closely as possible suggest a 25% overlap by chance alone.
Use of FDR for Union-Intersection tests

• Traditional
  – The ‘min’ test.
  – Low power
  – Not of definitive size
  – Ignores information (i.e., the p-value for min test is largest p-value for $h_0 \in H_0$ regardless of the value of any other p-values).

• Informational based approaches
  – All p-values are not equal
  – A variety of ways to weight
  – Let’s consider FDR or PTP – these are equal across datasets
  – Can conduct simple product of FDR.
Bioinformatics Issues

- HDB studies generate a huge amount of information.
- Storage and handling of the data can be difficult.
- Data standards are developing (MIAME for microarrays), proteomics just beginning.
End of Part 1
Statistical Analysis of Peptides
How to use MS for protein identification

Peptide mass fingerprinting

Example: peaks at m/z 333, 336, 406, 448, 462, 889
The only protein in the database that would produce these peaks is MALK|CGIR|GGSRPFLR|ATSK|ASR|SDD

- The exact protein needs to be in the database
- Works only with single protein fragmentations
Shotgun Protein Identification

Protein sample → protein identifications

Protein level
- enzymatic digestion

Peptide level
- Tandem mass spectrometry

MS/MS spectrum level
- MS/MS spectra

peptide grouping, validation
database search, validation

From Alexey Nesvizhskii
Example MS/MS spectrum
Interpretation of MS/MS data

• Direct interpretation ("de novo sequencing")
  – spectrum must be of good quality
  – the only identification method if the spectrum is not in the database
  – can give useful information (partial sequence) for database search

• General approach for database searching:
  – extract from the database all peptides that have the same mass as the precursor ion of the uninterpreted spectrum
  – compare each of them to the uninterpreted spectrum
  – select the peptide that is most likely to have produced the observed data

• MASCOT:
  – simple probabilistic model
  – calculate the probability that a peptide could have produced the given spectrum by chance
Threshold Model

- Threshold:
  - SEQUEST: $X_{corr} > 2.0$
  - $\Delta C_n > 0.1$
  - MASCOT: Ion Score $> 30$

- Sort by search score

- “Correct”
- Incorrect

[Image of a computer screen showing a spreadsheet with highlighted rows and columns, indicating correct and incorrect entries.]
Threshold Model: Bad Discrimination and Inconsistency

**Sensitivity:** fraction of all correct results passing filter

**Error Rate:** fraction of all results passing filter that are incorrect

*SEQUEST thresholds (from literature)*

*Test data (18 proteins): OMICS 6(2), 207 (2002)*

From Alexey Nesvizhskii
Difficulties in Interpreting Peptide Identifications based on MS/MS

Applies to both SEQUEST and Mascot (as it is used in practice) and, to large degree, to more recent tools

- No ‘useful’ measures of confidence
  (Mascot: ‘identity threshold’ guideline is not practical and rarely used)

- Different criteria used to filter data

- Unknown and variable false positive error rates

From Alexey Nesvizhskii
Just as assignment of quality scores to each base in DNA sequencing was essential for the genome sequencing programs, statistical models for estimating the accuracy of peptide and protein identifications are crucial for the success of high throughput proteomics.
Statistical Validation

- p-values or expectation values
  used, e.g., in sequence similarity searching

- Probabilities (Bayes)
  based on the ratio of two distributions (correct and incorrect) derived from the data (entire dataset)
  used, e.g., in information retrieval (relevant vs. non-relevant documents)
Expectation Values (empirical model)

\[ p(s_m) = \sum_{s \geq s_m} P(s) \]

probability to get score \( s \geq s_m \) by chance

\[ E(s_m) = N \sum_{s \geq s_m} P(s) \]

expected number of random matches with \( s \geq s_m \)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Peptide</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ISLLDAQSAPLR</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>VVEELCTPEGK</td>
<td>2.1</td>
</tr>
<tr>
<td>3</td>
<td>DLLLQWCWENCK</td>
<td>2.0</td>
</tr>
<tr>
<td>4</td>
<td>ECDVVSNTIIAEK</td>
<td>1.9</td>
</tr>
<tr>
<td>5</td>
<td>GDAVFVIDALNR</td>
<td>1.7</td>
</tr>
<tr>
<td>6</td>
<td>VPTPNVSVVTNR</td>
<td>1.6</td>
</tr>
<tr>
<td>7</td>
<td>SYLFCMGAEK</td>
<td>1.6</td>
</tr>
<tr>
<td>8</td>
<td>PEQSDLRSWTAK</td>
<td>1.5</td>
</tr>
</tbody>
</table>

... N peptides

From Alexey Nesvizhskii
Expectation Values (explicit model)

\[ P(s) = \frac{\mu^s}{s!} \exp(-\mu) \]

\( \mu \): function of mass tolerance, number of experimental peaks, number of calculated ions, mass, charge

Probability to get score \( s \geq s_m \) by chance

\[ p(s_m) = \sum_{s \geq s_m} P(s) \]

Geer et al. (upper bound)

Expected number of random matches with \( s \geq s_m \)

\[ E(s_m) = N(1 - (1 - \sum_{s \geq s_m} P(s))^N) \approx N^2 \sum_{s \geq s_m} P(s) \]

Sadygov et al. (lower bound)

From Alexey Nesvizhskii
Expectation Values (or p-values): Limitations

1. P-values or E-values are not well suited for the analysis of large-scale datasets (do not allow estimation of error rates as a function of filtering threshold)

   see, e.g., recent papers by Tsibshirani and others on the subject of p-values vs. False Discovery Rate (FDR) approach

2. Difficult to take advantage of other useful information (e.g., number of missed cleavages, peptide retention time)

3. Need to compute protein probabilities by combining probabilities of peptides corresponding to the same protein. Whether peptide expectation values can be used for that purpose is not clear

From Alexey Nesvizhskii
Modeling Large-Scale Datasets

entire dataset, \( M \) spectra
(1 or more LC/MS/MS runs)

Database

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>Peptide</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ISILDAQSAPIR</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>VVEELCTPEGK</td>
<td>3.9</td>
</tr>
<tr>
<td>3</td>
<td>DLLLQWCWENGK</td>
<td>1.2</td>
</tr>
<tr>
<td>4</td>
<td>ECDVVSNTIIAEK</td>
<td>0.9</td>
</tr>
<tr>
<td>5</td>
<td>GDAVFVIDALNR</td>
<td>3.6</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>SYLFCMFAEK</td>
<td>1.1</td>
</tr>
</tbody>
</table>

best match to each spectrum

raw score
E value
p-value

From Alexey Nesvizhskii
**Statistical Model for Computing Peptide Probabilities (PeptideProphet)**

**entire dataset:**

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>Peptide</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ISLIDAQSAPLR</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>VVEELCTPEGK</td>
<td>3.9</td>
</tr>
<tr>
<td>3</td>
<td>DLLLQWCWENCK</td>
<td>1.2</td>
</tr>
<tr>
<td>4</td>
<td>ECDVVSNTIIAEK</td>
<td>0.9</td>
</tr>
<tr>
<td>5</td>
<td>GDAVFVIDALNR</td>
<td>3.6</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>SYLFCMMEAEM</td>
<td>1.1</td>
</tr>
</tbody>
</table>

| spectrum | best match   | score  |


From Alexey Nesvizhskii
Statistical Model for Computing Peptide Probabilities (PeptideProphet)

<table>
<thead>
<tr>
<th>Spectrum Peptide Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ISLLDAQSAPLR         4.5</td>
</tr>
<tr>
<td>2 VVEELCTPEGK          3.9</td>
</tr>
<tr>
<td>3 DLLLQWCCWENGK        1.2</td>
</tr>
<tr>
<td>4 ECDVVSNTIIAEK        0.9</td>
</tr>
<tr>
<td>5 GDAVFVIDALNR         3.6</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>M SYLFCMEAEK           1.1</td>
</tr>
</tbody>
</table>

Entire dataset:

EM mixture model algorithm learns the most likely distributions among correct and incorrect peptide assignments given the observed data.

From Alexey Nesvizhskii
Illustration: Assigning Probabilities to Mascot Search Results

H. Influenzae, membrane fraction, 15 LC/MS/MS runs (~30,000 spectra)

From Alexey Nesvizhskii
Accuracy of Learned Distributions and Computed Probabilities

Database searched:
- Human
- H. Influenzae

Size ratio: ~ 20:1

For those familiar with "reverse database search" approach:
This is an equivalent of appending 20 randomized databases of equal size.

Method is accurate

H. Influenzae, membrane fraction, 15 LC/MS/MS runs
~30,000 spectra

From Alexey Nesvizhskii